

Article

Universal Approximation of Operators with Transformers and Neural Integral Operators

Emanuele Zappala *  and Maryam Bagherian 

Department of Mathematics and Statistics, Idaho State University, 921 S. 8th Ave., Stop 8085, Pocatello, ID 83209, USA; maryambagherian@isu.edu

* Correspondence: emanuelezappala@isu.edu

Abstract

We study the universal approximation properties of transformers and neural integral operators for operators in Banach spaces. In particular, we show that the transformer architecture is a universal approximator of integral operators between Hölder spaces. Moreover, we show that a generalized version of neural integral operators, based on the Gavurin integral, treats them as universal approximators of arbitrary operators between Banach spaces. Lastly, we show that a modified version of the transformer, which uses Leray–Schauder mappings, is a universal approximator of operators between arbitrary Banach spaces.

Keywords: operator learning; Hölder space; G -equivariant map; gavurin integral; leray-schauder mapping

MSC: 68T07; 47N40

1. Introduction

Operator learning constitutes a branch of deep learning dedicated to approximating nonlinear operators defined between Banach spaces. The appeal of this field stems from its capacity to model complex phenomena, such as dynamical systems, for which underlying governing equations remain elusive. Operator learning is thought to have originated with the theoretical groundwork established in [1] and subsequently implemented in [2].

Operator learning has been successfully applied in several fields of science, including learning neural integro-differential equations [3], integral operators [4,5], quantum state tomography [6], quantum circuit learning [7], control theory [8], partial differential equations [9], and holographic quantum chromodynamics [10], to mention but a few.

Transformers [11] have emerged as a powerful tool in natural language processing, and they have more recently been employed for addressing problems in operator learning, including dynamical systems. Previous research has explored their use in integral equations [4,5], inverse problems [12], fluid flows [13], and PDEs [14], among others.

Recent theoretical work highlights the strong approximation capabilities of transformer architectures across a range of settings. In [15], the authors showed that a variation in transformers' attention mechanism, which includes free parameters, allows the recovery of polynomials with zero error over a compact subset of a Euclidean space. This procedure has the drawback of potential exponential scaling of free parameters. It was shown in [16] that transformers can approximate sequence-to-sequence functions with respect to a suitable "average" error. The case of approximating sequence-to-sequence functions with some



Academic Editors: Jianjun Zhou and Jong-Min Kim

Received: 31 July 2026

Revised: 4 September 2026

Accepted: 8 September 2026

Published: 11 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

smoothness and with infinite dimensional inputs was studied in [17], where learnable positional encoding was employed. More recently, in [18], these works were further extended to show that sequence-to-sequence functions with Hölder and Sobolev regularity can be approximated in the L^p norm.

Recent studies have explored the application of transformers to operator modeling. For instance, [19] demonstrated the efficacy of attention mechanisms in learning operators with low regularity, specializing in Izhikevich and tempered fractional LIF neuron models. This work builds upon previous research [20,21] that established near-optimal error bounds for learning nonlinear operators between infinite-dimensional Banach spaces, notably DeepONets [22]. In addition, the work of [19] shows that on Banach spaces with the approximation property, transformers with positional encoding are universal approximators, using the addition of an encoder and a decoder and applying the results of [23] to uniformly approximate continuous functions via CNNs.

Transformers without positional encoders and without additional encoders and decoders have not yet been extensively explored in the case of operator learning problems. In fact, the results of [16–18], when applied directly to the proof of [19] in this setting, give error bounds that are either not uniform or are restricted to some classes of functions that are not directly related to the construction in [19].

The main results of the present paper are the following. We establish a universal approximation framework for Urysohn integral operators between Hölder spaces using transformer architectures without positional encoding, relying instead on their permutation-equivariant approximation properties. Our methods can also be applied for any equivariant universal approximator of sequence-to-sequence mappings, of which the transformer without positional encoding is an example. Moreover, incorporating Leray–Schauder mappings as encoders, we obtain a universal approximation result for (nonlinear) operators between arbitrary Banach spaces, including those without the approximation property. Lastly, we introduce a generalized neural integral operator based on the Gavurin integral and prove its universal approximation property for operators between Banach spaces under some smoothness assumption.

This paper is organized as follows. Section 2 outlines the necessary preliminaries. In Section 3, we prove that transformers are universal approximators for Hölder spaces. Section 4 investigates the universal approximation capabilities of neural integral operators for operators in Banach spaces. We conclude with Section 5, exploring further perspectives and open questions.

2. Preliminaries

2.1. Notation

Throughout this article, we assume a discretization of the interval $[a, b]$ as $a = t_0 < t_1 < \dots < t_{n-1} < t_n = b$. Unless otherwise explicitly stated, over any subinterval, i.e., $[a, t_1], \dots, [t_{n-1}, b]$, the point $x_i \in [t_i, t_{i+1}]$.

Let $X = C^{k,\alpha}([a, b])$ denote the Hölder space of k -differentiable functions with the k^{th} derivative that are Hölder continuous. We let $T : X \rightarrow X$ be an integral operator. Consider the values x_i and $z_k = y(t_k)$ for $i, k = 1, \dots, n$, where x_i and t_k vary among the points of the discretization determined by the integration approximation.

Let the kernel of T be the function $G : \mathbb{R} \times [a, b] \times [a, b] \rightarrow \mathbb{R}$, which is continuous with respect to all variables, and is continuously differentiable with respect to the first and third variables. Here, G takes as input a function u and two “time” variables from the interval $[a, b]$. Then, T is defined according to the assignment

$$T(u)(x) = \int_a^b G(u(t), x, t) dt.$$

In other words, T is an Urysohn integral operator.

For a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we indicate by $f^i(x_1, \dots, x_n)$ the i^{th} -entry of the vector $f(x_1, \dots, x_n)$.

Following the previous conventions, for each $i = 1, \dots, n$ in the rest of the paper, we indicate

$$T^i(u)(x_1, \dots, x_n) := \int G(u(t), x_i, t) dt,$$

where, as mentioned above, $x_i \in [t_i, t_{i+1}]$ for each i . So, $T(u)$ is the vector function whose i^{th} -entry function is $T^i(u)$. Given a function $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$, we introduce the following norm

$$\|F\|_p = \left(\sum_{i=1}^n \int_{[a, t_1] \times \dots \times [t_{n-1}, b]} |F^i(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}}, \quad (1)$$

which corresponds to the Euclidean norm of the component-wise p -norm, and can be applied to $T^i(u)$ as a particular case. This p -norm will be the objective of the approximation results over Hölder spaces for transformers.

2.2. Leray–Schauder Mappings

Leray–Schauder mappings [24,25] are a fundamental tool in nonlinear functional analysis. They are used to establish the existence of solutions to nonlinear problems by extending the concept of degrees from finite-dimensional spaces to infinite-dimensional Banach spaces. Roughly speaking, given a compact set K and a finite dimensional subspace of a Banach space, we want to find a continuous mapping onto the finite subspace satisfying the property that the norm of elements in the compact set and the image of the mapping is smaller than a fixed $\epsilon > 0$. Following [25], we define Leray–Schauder mappings on a Banach space X as follows. Let K be a compact set, and E_n an n -dimensional space spanned by the ϵ -net $x_1, \dots, x_n$ of K . We define $P_n : X \rightarrow E_n$ by the assignment

$$P_n x = \frac{\sum_{i=1}^n \mu_i(x) x_i}{\sum_{i=1}^n \mu_i(x)},$$

where

$$\mu_i(x) = \begin{cases} \epsilon - \|x - x_i\|, & \|x - x_i\| \leq \epsilon \\ 0, & \|x - x_i\| > \epsilon \end{cases} \quad (2)$$

for all $i = 1, \dots, n$. One sees directly that this mapping satisfies the property that

$$\|x - P(x)\| < \epsilon,$$

for all $x \in K$.

2.3. Gavurin Neural Integral Operator $\mathfrak{T}(x)$

We hereby recall the definition of the Gavurin integral [26], and use this to define a class of neural integral operators of relevance for the rest of this paper.

We follow [27], Chapter 17, and consider a mapping $R : [x_0, \bar{x}] \subset X \rightarrow B(X, Y)$, where $B(X, Y)$ denotes the space of bounded linear operators from X to Y , two Banach spaces.

Let $0 = t_0 < t_1 < \dots < t_n = 1$ be a partition of $[0, 1]$, and let $\tau_k \in [t_k, t_{k+1}]$. Then, the integral of R is defined in a procedure similar to the classical Riemann integral as

$$\int_{x_0}^{x_0+\Delta x} R(x)dx = \int_0^1 R(x_0 + t\Delta x)\Delta x dt = \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} R(x_0 + \tau_k \Delta x)\Delta x (t_{k+1} - t_k), \quad (3)$$

when the limit exists and is independent of the choices of τ_k . We indicate the action of the integral operator on an element z of X as $\int_{x_0}^{x_0+\Delta x} R(x)(z)dx$.

We define a *Gavurin integral operator* to be an integral operator $T : X \rightarrow Y$, which is the point-wise convex sum of the integral operators, as in Equation (3), as follows:

$$T(z) = \sum_i \eta_i(z) [F_i(z) + \int_{x_0}^{x_0+\Delta x} R_i(x)(z)dx], \quad (4)$$

where for every z , $\sum_i \eta_i(z) = 1$, i.e., the coefficients are a partition of unity, and F_i values are arbitrary linear operators.

We define a *Gavurin neural integral operator* to be an operator approximating the form in (4) with a Riemann sum consisting of Leray–Schauder maps and neural networks in the following sense. Let P and P' denote two Leray–Schauder mappings, $P : U \rightarrow X$ and $P' : U' \rightarrow Y$, where U and U' are open sets containing two given compacts, $\mathbb{K}$ and $\mathbb{K}'$. Then, $P(\mathbb{K})$ and $P'(\mathbb{K}')$ both span the finite dimensional subspaces E and E' of X and Y , respectively, by construction. We let φ and φ' denote two linear homeomorphisms $\varphi : E \rightarrow \mathbb{R}^n$ and $\varphi' : E' \rightarrow \mathbb{R}^m$, where we have set $\dim E = n$ and $\dim E' = m$. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a neural network, and consider the composition $F_{P,f} := \varphi'^{-1} f \varphi P$. Then, a Gavurin neural integral operator is a neural network defined as

$$F_{P,f_1,\dots,f_q,g_1,\dots,g_q}(z) := \sum_i \eta_i(z) F_{P,f_i}(z) + \sum_r \sum_i \sum_k \eta_i(z) \alpha_k(t_r) F_{P,g_i}(z),$$

where $\{\eta_i\}_i$ is a partition of unity subordinate to an open cover of $\mathbb{K}$, $\{\alpha_k\}_k$ is a partition of unity of $[0, 1]$, and $0 = t_0 \leq \dots \leq t_q = 1$ is a discretization of the interval $[0, 1]$. All sums in the definition are finite, and $F_{P,f_1,\dots,f_q,g_1,\dots,g_q}$ implicitly depends on the chosen discretization of $[0, 1]$, as well as the partitions of unity $\{\eta_i\}_i$ and $\{\alpha_k\}_k$. We notice that the sum over r is substantially mimicking a Riemann sum, therefore approximating an integral.

2.4. Models with Permutation-Equivariance

Equivariance can be formalized using the concept of a G -set for a group G consisting of a set S and where G acts on S from the right. Given two G -sets, X and Y , for the same group G , a function $f : X \rightarrow Y$ is said to be G -equivariant if $f(x \cdot g) = f(x) \cdot g$ for all $g \in G$ and all $x \in X$.

A *transformer* is a sequence-to-sequence model composed of an encoder and a decoder, both built from stacks of self-attention and feedforward layers. A transformer defines a function $T : \mathcal{X}^n \rightarrow \mathcal{Y}^m$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and output token spaces, respectively, and n, m are the input and output sequence lengths. Given an input sequence of tokens $x = (x_1, \dots, x_n)$, each token is mapped to a vector via an embedding function $\mathbf{e}_i = \text{Embed}(x_i) \in \mathbb{R}^d$. A positional encoding $\mathbf{p}_i \in \mathbb{R}^d$ is added to retain order information $\mathbf{z}_i^{(0)} = \mathbf{e}_i + \mathbf{p}_i$. At each layer ℓ , the input $\mathbf{Z}^{(\ell-1)} = (\mathbf{z}_1^{(\ell-1)}, \dots, \mathbf{z}_n^{(\ell-1)})$ is transformed via multi-head self-attention, as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where d_k is the length of the key vectors, used as a scaling factor to keep numbers stable,

$$\mathbf{z}_i^{(\ell, \text{attn})} = \text{MultiHead}(\mathbf{z}_i^{(\ell-1)}) = \text{Concat}(h_1, \dots, h_H)W^O,$$

where each head h_j is

$$h_j = \text{Attention}(Q_j, K_j, V_j), \quad Q_j = \mathbf{Z}^{(\ell-1)}W_j^Q, \quad K_j = \mathbf{Z}^{(\ell-1)}W_j^K, \quad V_j = \mathbf{Z}^{(\ell-1)}W_j^V.$$

Each position i is updated via a position-wise feedforward network:

$$\mathbf{z}_i^{(\ell, \text{ff})} = \text{FFN}(\mathbf{z}_i^{(\ell, \text{attn})}) = \sigma(\mathbf{z}_i W_1 + b_1)W_2 + b_2.$$

The full update rule at each layer ℓ is

$$\begin{aligned} \mathbf{z}_i^{(\ell, \text{norm1})} &= \text{LayerNorm}(\mathbf{z}_i^{(\ell-1)} + \mathbf{z}_i^{(\ell, \text{attn})}), \\ \mathbf{z}_i^{(\ell)} &= \text{LayerNorm}(\mathbf{z}_i^{(\ell, \text{norm1})} + \mathbf{z}_i^{(\ell, \text{ff})}). \end{aligned}$$

In the context of transformers, particularly in machine learning models, *equivariance* refers to equivariance with respect to the natural symmetric group action on sequences of inputs–outputs. More specifically, for the function $f : X \rightarrow Y$ (e.g., a neural network layer), and the group G acting on both X and Y , the function f is said to be equivariant with respect to G if

$$f(x \cdot g) = f(x) \cdot g, \quad \forall g \in G, x \in X,$$

where $x \cdot g$ and $f(x) \cdot g$ denote the actions of g on X and Y , respectively.

Deep Sets [28] are explicitly designed to be permutation-invariant or -equivariant by applying a shared function to each element, followed by a symmetric aggregation operation (e.g., summation). Neural Functional models [29] generalize this idea by treating input data as functions over domains and ensure equivariance under transformations such as permutations by construction. Transformers [11], before the addition of positional encodings, are naturally permutation-equivariant due to the symmetric structure of the self-attention mechanism, i.e., each token tends to all others through shared attention weights, making the output invariant to the order of inputs under shared weights. However, once positional information is introduced, this permutation symmetry is typically broken.

The approach pursued in this article can be applied to any architecture that is equivariant and a universal approximator of sequence-to-sequence mappings, in the norm defined in (1). In other words, given an approximator of sequence-to-sequence mappings, we can show that there is a corresponding construction that is a universal approximator of nonlinear operators between the Hölder spaces. We pose the following useful definition.

In this article, we will use the term *generalized transformer* to denote a mapping $\mathcal{W}$ defined on a space of functions as follows. We have a locally finite partition of unity $\{\eta_i\}$ defined on the space subordinate to some open covering $\mathcal{U}$, and we let W_U denote a standard transformer for each open U in the covering $\mathcal{U}$, whether it has positional encoding or not. For a function $f \in \mathcal{U}$, $W_U(f)$ is defined by applying f to some pre-defined sampling scheme of the domain of the space of functions, and then applying W_U on the corresponding outputs of f . We then set

$$\mathcal{W}(f) = \sum_i \eta_i(f)W_U(f).$$

Each sum is finite since the partition of unity is locally finite. Details of this procedure are provided below in Section 3.

2.5. Notation and Generalities

The notations used in the following sections are summarized in Table 1. The precise definitions of each notation are given in the main text and are used throughout this paper.

Table 1. Summary of notations used in the rest of the manuscript.

Notation	Description
T	The mapping between Banach spaces to be approximated
$\mathcal{T}$	The integration scheme approximating an integral operator T
$T^i(u)$	i^{th} -component of the function induced by evaluating the operator T at u
$\mathfrak{T}$	Leray–Schauder transformer or Gavurin neural integral operator
W_y	Transformer architecture associated to a function y
$\mathcal{W}$	Transformer architecture obtained via partition of unity
F	A real vector function $F : \mathbb{R}^n \times [a, b]^n \rightarrow \mathbb{R}^n$
F_y	Function $[a, b]^n \rightarrow \mathbb{R}^n$ induced by F and y
F_y^i	i^{th} -component of F_y

In this work, the term *operator* will refer to any continuous mapping, without restriction of linearity. This follows the common notations in operator learning, where nonlinear maps between function spaces are also called operators.

3. Universal Approximation for Transformers

We recall the following definitions. Let $f : [a, b] \rightarrow \mathbb{R}$ be a function. Then, we say that f is *Hölder continuous* if there exists a constant c such that

$$|f(x) - f(y)| \leq c|x - y|^\alpha,$$

for all $x, y \in [a, b]$, where $\alpha \in (0, 1]$. We define the set $C^{k,\alpha}([a, b])$ of all Hölder functions whose k th derivative is Hölder continuous, and define the norm

$$\|f\|_{k,\alpha} = \|f\|_k + \sup\left\{\frac{|f^k(x) - f^k(y)|}{|x - y|^\alpha} \mid x, y \in [a, b], x \neq y\right\},$$

where $\|f\|_k := \max_{r \leq k} \|f^r\|_\infty$ is the maximum among the uniform norms of the derivatives of f . The space $C^{k,\alpha}([a, b])$ is a Banach space. In this section, we consider $C^{k,\alpha}([a, b])$, where $1 \leq k \in \mathbb{N}$ and $\alpha \in (0, 1]$, and let $T : C^{k,\alpha}([a, b]) \rightarrow C^{k,\alpha}([a, b])$ be a given integral operator. We want to show that there is a transformer architecture $\mathcal{W}$ that approximates T with arbitrarily high precision in the following sense. For an arbitrary choice of $\epsilon > 0$, there exist a discretization $a = t_0 < t_1 < \dots < t_{n-1} < t_n = b$ of $[a, b]$ and a transformer architecture $\mathcal{W}$ such that for each $u \in C^{k,\alpha}([a, b])$, u in some fixed set, we have that

$$\left(\sum_{i=1}^n \int_{[a,t_1] \times \dots \times [t_{n-1},b]} |T_u^i(x_1, \dots, x_n) - \mathcal{W}_u^i(x_1, \dots, x_n)|^p d\mu\right)^{\frac{1}{p}} < \epsilon, \quad (5)$$

where $d\mu$ is a shorthand notation for $dx_1 \dots dx_n$.

We will need several lemmas before obtaining the main result.

We set X to be a Hölder space $C^{k,\alpha}([a, b])$ (with $k \geq 1$) of functions defined on some closed interval $[a, b]$. Let $T : X \rightarrow X$ denote an integral operator, where we assume that T is an Urysohn integral operator and that the kernel of T is a function $G : \mathbb{R} \times [a, b] \times [a, b] \rightarrow \mathbb{R}$ that is continuous with respect to all variables and is continuously differentiable with respect to the first and third variables. We let $\mathbb{B}_\rho$ be a ball centered at the origin, with

radius ρ fixed. In this article, we will consider only numerical integration schemes that are non-adaptive and that are given by continuous formulas with respect to the sampled points.

Lemma 1. *Let T and X be as above. Then, we can find an integration scheme, $\mathcal{T}$, such that*

$$\|T(y) - \mathcal{T}(y)\|_{\infty} < \epsilon, \quad (6)$$

for all $y \in \mathbb{B}_{\rho}$, where $\|\cdot\|_{\infty}$ denotes the uniform norm.

Proof. We denote by $G : \mathbb{R} \times [a, b] \times [a, b] \rightarrow \mathbb{R}$ the kernel of the integral operator T . Therefore,

$$T(y)(x) = \int_a^b G(y(t), x, t) dt,$$

for all $x \in [a, b]$. For $y \in \mathbb{B}_{\rho}$, from $\|y\| \leq \rho$, it follows that $\sup_{[a, b]} |y(x)| \leq \rho$. Therefore, $G(y(t), x, t)$ takes values $y(t)$ in $[-\rho, \rho]$. Let M_1 denote the value

$$M_1 := \max_{[-\rho, \rho] \times [a, b] \times [a, b]} |\partial_y G(y(t), x, t)|,$$

and let

$$M_2 := \max_{[-\rho, \rho] \times [a, b] \times [a, b]} |\partial_t G(y(t), x, t)|.$$

Let us choose a quadrature rule with n points $\{q_i\}_{i=1}^n$ selected from $[a, b]$ to approximate the integral operator T . We indicate the quadrature by $\mathcal{T}$: $T(y)(x) = \int_a^b G(y(t), x, t) dt \approx \sum_{i=1}^n G(y(q_i), x, q_i) w_i =: \mathcal{T}(y)(x)$, where the coefficients w_i are the weights of the quadrature, and $G(y(q_i), x, q_i)$ is continuous with respect to x . Also, $G(y(t), x, t)$ is differentiable with respect to the variable t of an integration with a derivative given by

$$\frac{dG(y(t), x, t)}{dt} = \partial_t G(y(t), x, t) + \partial_y G(y(t), x, t) \cdot y'(t),$$

for any chosen $y \in \mathbb{B}_{\rho}$. For the sake of simplicity, we use the forward rectangle method in what follows, but a similar procedure (upon suitably changing the error bounds and assuming that the index k of the Hölder space is large enough) can be performed for different integration rules, such as the Midpoint, Trapezoidal, corrected Trapezoidal, Cavalieri–Simpson, Boole, etc. It is known that the error in performing the numerical integration of $\int_a^b g(t) dt$ is bounded by $\frac{(b-a)^2}{2n} \sup_{[a, b]} |g'(t)|$ [30]. Therefore, for each fixed $x_0 \in [a, b]$, we have the bound

$$\begin{aligned} |T(y)(x_0) - \mathcal{T}(y)(x_0)| &\leq \frac{(b-a)^2}{2n} \sup_{[a, b]} |\partial_t G(y(t), x_0, t) + \partial_y G(y(t), x_0, t) \cdot y'(t)| \\ &\leq \frac{(b-a)^2}{2n} [\sup_{[a, b]} |\partial_t G(y(t), x_0, t)| + \sup_{[a, b]} |\partial_y G(y(t), x_0, t) \cdot y'(t)|] \\ &\leq \frac{(b-a)^2}{2n} (M_1 \rho + M_2), \end{aligned}$$

from which we derive $|T(y)(x) - \mathcal{T}(y)(x)| \leq \frac{(b-a)^2}{2n} (M_1 \rho + M_2)$ for all $y \in \mathbb{B}_{\rho}$ and all $x \in [a, b]$. Upon choosing $n > (b-a)^2 \frac{M_1 \rho + M_2}{2\epsilon}$, we find that $\|T(y) - \mathcal{T}(y)\|_{\infty} < \epsilon$ uniformly in $\mathbb{B}_{\rho}$. $\square$

A variation in the following argument appears also in [4,5]. We provide here a more formal approach. Before stating and proving the result, we pose the following definition.

Definition 1. Let σ denote a permutation of n elements, i.e., $\sigma \in \Sigma_n$, where Σ_n is the permutation group on n elements. Then, σ acts on the tuples of type $(z_1, \dots, z_n, x_1, \dots, x_n) \in \mathbb{R}^{2n}$ through the induced diagonal permutation representation, which just reorders the entries $z_1, \dots, z_n$ and $x_1, \dots, x_n$ according to the permutation σ , as follows:

$$(z_1, \dots, z_n, x_1, \dots, x_n) \cdot_1 \sigma := (z_{\sigma(1)}, \dots, z_{\sigma(n)}, x_{\sigma(1)}, \dots, x_{\sigma(n)}).$$

We indicate by $\cdot_2$ the permutation action of Σ_n on n -tuples as

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \cdot_2 \sigma = \begin{bmatrix} x_{\sigma(1)} \\ \vdots \\ x_{\sigma(n)} \end{bmatrix}.$$

We will also denote by $\cdot_2$ the permutation action on tuples such as $(x_1, \dots, x_n)$ to shorten the notation.

Lemma 2. The integration scheme from Lemma 1 induces a continuous function $F : \mathbb{R}^n \times [a, b]^n \rightarrow \mathbb{R}^n$. Moreover, F is equivariant with respect to the actions defined in Definition 1.

Proof. We set $F^i(z_1, \dots, z_n, x_1, \dots, x_n) := \sum_{k=1}^n G(z_k, x_i, t_k) w_k$ for each $i = 1, \dots, n$, where the points $t_1, \dots, t_n$ are determined by the integration scheme defined in Lemma 1. We can therefore define F as the column vector $[F^i]_{i=1}^n$ and as a function of the tuple $(z_1, \dots, z_n, x_1, \dots, x_n)$. Since each function F^i is continuous with respect to its entries z_k and x_i , it follows that the function F is also continuous.

To show that the function F just defined is permutation-equivariant, using the same notations introduced above, we compute

$$\begin{aligned} F((z_1, \dots, z_n, x_1, \dots, x_n) \cdot_1 \sigma) &= F(z_{\sigma(1)}, \dots, z_{\sigma(n)}, x_{\sigma(1)}, \dots, x_{\sigma(n)}) \\ &= \begin{bmatrix} F^{\sigma(1)}(z_{\sigma(1)}, \dots, z_{\sigma(n)}, x_1, \dots, x_n) \\ \vdots \\ F^{\sigma(n)}(z_{\sigma(1)}, \dots, z_{\sigma(n)}, x_1, \dots, x_n) \end{bmatrix} \\ &= \begin{bmatrix} \sum_{k=1}^n G(z_{\sigma(k)}, x_{\sigma(1)}, t_{\sigma(k)}) w_{\sigma(k)} \\ \vdots \\ \sum_{k=1}^n G(z_{\sigma(k)}, x_{\sigma(n)}, t_{\sigma(k)}) w_{\sigma(k)} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{k=1}^n G(z_k, x_{\sigma(1)}, t_k) w_k \\ \vdots \\ \sum_{k=1}^n G(z_k, x_{\sigma(n)}, t_k) w_k \end{bmatrix} \\ &= \begin{bmatrix} F^{\sigma(1)}(z_1, \dots, z_n, x_1, \dots, x_n) \\ \vdots \\ F^{\sigma(n)}(z_1, \dots, z_n, x_1, \dots, x_n) \end{bmatrix} \\ &= F(z_1, \dots, z_n, x_1, \dots, x_n) \cdot_2 \sigma. \end{aligned}$$

This shows that F is equivariant with respect to the actions. $\square$

In fact, we will only need a subcase of the previous lemma. We formulate this in the following result.

Lemma 3. For any choice of function y in the Hölder space $C^{k,\alpha}([a, b])$, the integration scheme from Lemma 1 induces a permutation-equivariant continuous function $F_y : [a, b]^n \rightarrow \mathbb{R}^n$.

Proof. For a fixed $y \in C^{k,\alpha}([a, b])$, the function F constructed in Lemma 2 will have the first n entries only depending on t_i , since they are obtained by evaluating y at the points t_i , for $i = 1, \dots, n$. The points t_i , however, are determined by the integration scheme and are therefore fixed. This induces a function $F_y : [a, b]^n \rightarrow \mathbb{R}^n$, which is continuous. To show that the map F_y is permutation-equivariant with respect to the $\cdot_2$ action, we use the equivariance of F proved in Lemma 2 in the following way. First, observe that F is independent of any permutation with respect to the first n coordinates, since they are all summed together. The diagonal permutation action σ on the entries of $F(y_1, \dots, y_n, x_1, \dots, x_n)$ is the same as the action of σ on the entries of $F_y(x_1, \dots, x_n)$ for $y_1, \dots, y_n$, fixed as $y_r = y(t_r)$. Therefore, the equivariance of Lemma 2 completes the proof. $\square$

Recall the following notation conventions, which will be used in the subsequent results. For a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we indicate by $f^i(x_1, \dots, x_n)$ the i^{th} entry of the vector $f(x_1, \dots, x_n)$. Following Lemma 2, we also define the function F , whose i^{th} component is given by $F^i(z_1, \dots, z_n, x_1, \dots, x_n) := \sum_{k=1}^n G(z_k, x_i, t_k) w_k$ for each $i = 1, \dots, n$, and F_y is induced from F as in Lemma 3. We notice that F_y gives an approximation of $T^i(y)$ by construction.

Lemma 4. Let y be in the ball $\mathbb{B}_\rho$ of radius ρ in the Hölder space $C^{k,\alpha}([a, b])$. Then, for any choice of $\epsilon > 0$, we can find a transformer $W_y : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that

$$\left(\sum_{i=1}^n \int_{[a,t_1] \times \dots \times [t_{n-1},b]} |F_y^i(x_1, \dots, x_n) - W_y^i(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}} < \epsilon. \quad (7)$$

Over a compact $\mathbb{K} \subset C^{k,\alpha}([a, b])$, this approximation can be conducted uniformly in y with a generalized transformer $\mathcal{W}$, as follows:

$$\left(\sum_{i=1}^n \int_{[a,t_1] \times \dots \times [t_{n-1},b]} |F_y^i(x_1, \dots, x_n) - \mathcal{W}^i(y)(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}} < \epsilon. \quad (8)$$

Proof. Applying Lemma 3, for any choice of $y \in \mathbb{B}_\rho$, $\mathcal{T}$ induces a continuous and permutation-equivariant map $F_y : [a, b]^n \rightarrow \mathbb{R}^n$. Applying Theorem 2 of [16] to the permutation-equivariant map F_y , we can find a transformer architecture W_y , which depends on the given y , such that $\|F_y - W_y\|_p < \epsilon$. Then, we have the following:

$$\begin{aligned} \|F_y - W_y\|_p &= \left(\sum_{i=1}^n \int_{[a,b]^n} |F_y^i(x_1, \dots, x_n) - W_y^i(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}} \\ &\geq \left(\sum_{i=1}^n \int_{[a,t_1] \times \dots \times [t_{n-1},b]} |F_y^i(x_1, \dots, x_n) - W_y^i(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}}. \end{aligned}$$

We therefore obtain

$$\left(\sum_{i=1}^n \int_{[a,t_1] \times \dots \times [t_{n-1},b]} |F_y^i(x_1, \dots, x_n) - W_y^i(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}} < \epsilon.$$

For the second part of the statement, suppose that $\mathbb{K}$ is compact. Then, by the boundedness of the compacts, there exists a ball $\mathbb{B}_\rho$ of radius $\rho > 0$ that contains $\mathbb{K}$. For any choice of $y \in \mathbb{K} \subset \mathbb{B}_\rho$, we know from the first part of the lemma that $\mathcal{T}(y)$ is approximated arbitrarily well by some transformer W_y .

Observe that $\mathcal{T}(y)$ depends continuously on y by construction (see Lemma 1). Moreover, as discussed in Lemma 1, in $\mathbb{B}_\rho$, the ranges of the functions y are contained in $[-\rho, \rho]$. Using the Heine–Cantor Theorem for G over the interval $[-\rho, \rho]$ and the definition of F_y (given in Lemma 2) induced by $\mathcal{T}(y)$, we can see that upon choosing an $r > 0$ value that is small enough for any y, z such that $\|y - z\|_{k,\alpha} < r$, we have the inequality

$$\left(\sum_i \int_{[a,t_1] \times \dots \times [t_{n-1},b]} |F_y^i(x_1, \dots, x_n) - F_z^i(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}} < \frac{\epsilon}{2}. \quad (9)$$

We choose an r -net for the compact $\mathbb{K}$ consisting of finitely many points $y_1, \dots, y_n$, with corresponding balls $\mathbb{B}_j$ with radii less than r . For each y_j , we find a transformer W_j such that

$$\left(\sum_{i=1}^n \int_{[a,t_1] \times \dots \times [t_{n-1},b]} |F_j^i(x_1, \dots, x_n) - W_j^i(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}} < \frac{\epsilon}{2},$$

where we use j in the subscript of F^i instead of y_j for simplicity. There exists a partition of unity $\{\eta_k\}$ subordinate to the covering $\{\mathbb{B}_k\}$. We use such partition of unity to define a transformer $\mathcal{W}$ as a combination of the W_k as

$$\mathcal{W}(y) = \sum_k \eta_k(y) W_k.$$

For any y in $\mathbb{K}$, making use of Jensen's inequality, we have

$$\begin{aligned} & \sum_{i=1}^n \int_{\Omega} |F_y^i(x_1, \dots, x_n) - \mathcal{W}^i(y)(x_1, \dots, x_n)|^p d\mu \\ &= \sum_{i=1}^n \int_{\Omega} \left| \sum_k \eta_k(y) F_y^i(x_1, \dots, x_n) - \sum_k \eta_k(y) W_k^i(x_1, \dots, x_n) \right|^p d\mu \\ &\leq \sum_{i=1}^n \sum_k \eta_k(y) \int_{\Omega} |F_y^i(x_1, \dots, x_n) - W_k^i(x_1, \dots, x_n)|^p d\mu \\ &\leq \sum_{i=1}^n \left[2^{p-1} \sum_k \eta_k(y) \int_{\Omega} |F_y^i(x_1, \dots, x_n) - F_k^i(x_1, \dots, x_n)|^p d\mu \right. \\ &\quad \left. + 2^{p-1} \sum_k \eta_k(y) \int_{\Omega} |F_k^i(x_1, \dots, x_n) - W_k^i(x_1, \dots, x_n)|^p d\mu \right] \\ &= 2^{p-1} \sum_k \eta_k(y) \sum_{i=1}^n \int_{\Omega} |F_y^i(x_1, \dots, x_n) - F_k^i(x_1, \dots, x_n)|^p d\mu \\ &\quad + 2^{p-1} \sum_k \eta_k(y) \sum_{i=1}^n \int_{\Omega} |F_k^i(x_1, \dots, x_n) - W_k^i(x_1, \dots, x_n)|^p d\mu \\ &< 2^{p-1} \sum_k \eta_k(y) \frac{\epsilon^p}{2^p} + 2^{p-1} \sum_k \eta_k(y) \frac{\epsilon^p}{2^p} \\ &= \frac{\epsilon^p}{2} + \frac{\epsilon^p}{2} \\ &= \epsilon^p, \end{aligned}$$

where $\Omega = [a, t_1] \times \dots \times [t_{n-1}, b]$. Taking p -roots, we complete the proof. $\square$

We consider the same class of maps as in the discussion preceding Lemma 1. Namely, we let $T : X \rightarrow X$ denote an Urysohn integral operator, such that the kernel of T is a function $G : \mathbb{R} \times [a, b] \times [a, b] \rightarrow \mathbb{R}$ that is continuous with respect to all variables and is

continuously differentiable with respect to the first and third variables. We are now in a position to prove the main result of this article.

Theorem 1. *Let $T : C^{k,\alpha}([a, b]) \rightarrow C^{k,\alpha}([a, b])$ be an integral operator as above, and let $\mathbb{K} \subset C^{k,\alpha}([a, b])$ be a compact set. Then, for any $\epsilon > 0$ and every choice of $p \geq 1$, there exists a generalized transformer architecture $\mathcal{W}$ and a discretization of $[a, b]$ (which depend on $\mathbb{K}$, ϵ and p) such that*

$$\left(\sum_i \int_{\Omega} |T^i(y)(x_1, \dots, x_n) - \mathcal{W}^i(y)(x_1, \dots, x_n)|^p \right)^{\frac{1}{p}} < \epsilon,$$

for every $y \in \mathbb{K}$, where $\Omega = [a, t_1] \times \dots \times [t_{n-1}, b]$ is determined by the discretization of $[a, b]$.

Proof. Without loss of generality, we set $[a, b] = [0, 1]$. We let $\bar{B}_\rho$ denote a ball containing $\mathbb{K}$. Applying Lemma 1, we find an integration scheme $\mathcal{T}$ such that

$$\|T(y) - \mathcal{T}(y)\|_\infty < \frac{\epsilon}{2},$$

with some discretization $0 = t_1, \dots, t_n = 1$. We can assume that the discretization is uniform, with subintervals of length $\frac{1}{n}$, since we can otherwise pass to a finer partition with this property, with $\mathcal{T}$ still satisfying the previous estimate. Using Lemma 3, we find that $\mathcal{T}$ induces a continuous permutation-equivariant map, which we still denote by $\mathcal{T}$ by an abuse of notation. We apply Lemma 4 as follows. Over the compact $\mathbb{K}$, there exists a generalized transformer $\mathcal{W}$ such that

$$\left(\sum_{i=1}^n \int_{\Omega} |\mathcal{T}^i(y)(x_1, \dots, x_n) - \mathcal{W}^i(y)(x_1, \dots, x_n)|^p d\mu \right)^{\frac{1}{p}} < \frac{\epsilon}{2},$$

where we have set again for simplicity $\Omega = [0, t_1] \times \dots \times [t_{n-1}, 1]$ to shorten notation, where t_i are the points of the aforementioned discretization.

Therefore, we have

$$\begin{aligned} & \sum_i \int_{\Omega} |T^i(y)(x_1, \dots, x_n) - \mathcal{W}^i(y)(x_1, \dots, x_n)|^p d\mu \\ & \leq \sum_{i=1}^n 2^{p-1} \int_{\Omega} |T^i(y)(x_1, \dots, x_n) - \mathcal{T}^i(y)(x_1, \dots, x_n)|^p d\mu \\ & \quad + \sum_{i=1}^n 2^{p-1} \int_{\Omega} |\mathcal{T}^i(y)(x_1, \dots, x_n) - \mathcal{W}^i(y)(x_1, \dots, x_n)|^p d\mu \\ & < \sum_{i=1}^n (2^{p-1} \frac{\epsilon^p}{2^p n^n} + 2^{p-1} \frac{\epsilon^p}{2^p n^n}) \\ & \leq \frac{\epsilon^p}{2} + \frac{\epsilon^p}{2} \\ & = \epsilon^p \end{aligned}$$

for all y in $\mathbb{K}$. Taking p -roots completes the proof. $\square$

Remark 1. We note that no special property in the proofs above is used regarding the architecture of the transformer, except the fact that transformers can approximate equivariant sequence-to-sequence maps in the norm (1) with arbitrary accuracy. As a consequence, the approach of this article can be used to show that any universal approximator of equivariant sequence-to-sequence mappings can be used to approximate integral operators in Hölder space with arbitrary precision. Our main

focus on transformers only stems from the fact that they are widely used in the study of operator learning problems.

Remark 2. The approximation result of Theorem 1 refers to a p -norm that is an average over the points where the tokens for the transformer can be sampled. While the error “on average” is small, for the learned operator, this does not mean that certain choices of values cannot incur a significant spike in the error. This type of behavior has in fact been observed concretely in [31], where such error spikes were avoided by training the transformer model in a Sobolev space that functioned for regularization.

The construction of the partition of unity over a compact subset of a normed space can be achieved via the same procedure that produces Leray–Schauder mappings, and it is therefore implementable numerically in practice. In fact, by fixing $\epsilon > 0$, we can cover a compact via finitely many balls centered at functions f_i . With μ_i defined as in (2), setting $\eta_i(x) = \frac{\mu_i(x)}{\sum_j \mu_j(x)}$ then gives a partition of unity subordinate to the initial covering. The functions f_i used to cover the dataset space can be learned by approximating them via neural networks.

Example 1. Here, we provide a discussion regarding possible applications of Theorem 1, along with some classes of problems that can be treated with these methods. A detailed study related to these considerations would be desirable.

Let $\mathbb{K}$ be a compact subset of $C^{k,\alpha}([a, b])$. The metric entropy, or Kolmogorov ϵ -entropy, measures the number of balls of radius $\epsilon > 0$ needed to cover $\mathbb{K}$. Therefore, we can estimate the number of functions in the partitions of unity used in the proofs based on the metric properties of $\mathbb{K}$ and the chosen accuracy. Once the partitions of unity have been fixed, we can approximate the target operator by a combination of partitions of unity and sequence-to-sequence maps through the discretization of the integration space, as in the proof of Theorem 1 and its preliminary lemmas. The results and methods of [17] give some asymptotic estimates on the parameters needed to approximate sequence-to-sequence functions, and therefore, combining the two approaches, it is possible to attempt an a priori study of the size of the architectures and number of transformers needed to approximate a target operator.

Classes of examples of problems that are studied via integral operators of the Urysohn type with sufficiently smooth kernels are nonlocal problems such as, for example, the Amari [32,33] and Wilson–Cowan [34] models in computational neuroscience. These models give rise to continuous maps between Hölder spaces [35] due to the regularity of the Urysohn kernels. Therefore, it is in principle possible to estimate, via the procedure described above, the size of the transformer architectures needed to approximate the operators on given compact subsets of $C^{k,\alpha}([a, b])$. More generally, we expect that a similar approach can be applied to problems described via Volterra and Fredholm integral equations, or PDEs with sufficiently regular nonlocal interactions.

While this is not the main scope of this article, we provide another universal approximation result for transformer architectures that implement a Leray–Schauder mapping, similar to [36]. For this purpose, we define a Leray–Schauder transformer to consist of a Leray–Schauder mapping composed of either a traditional transformer architecture with positional encoding as in [16,23] or the formulation of attention found in [15]. Let P and P' denote two Leray–Schauder mappings $P : U \rightarrow X$ and $P' : U' \rightarrow Y$, where U and U' are open sets containing two given compacts $\mathbb{K}$ and $\mathbb{K}'$. Then, $P(\mathbb{K})$ and $P'(\mathbb{K}')$ both span finite dimensional subspaces E and E' of X and Y , respectively, by construction. We let φ and φ' denote two linear homeomorphisms $\varphi : E \rightarrow \mathbb{R}^n$ and $\varphi' : E' \rightarrow \mathbb{R}^m$, where we set $\dim E = n$ and $\dim E' = m$. Let us indicate by W a transformer architecture as in [15]

or [16,23], where $W : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Then, a Leray–Schauder transformer $\mathfrak{T}$ is defined by the composition $\varphi'^{-1}W\varphi P : U \subset X \rightarrow E' \subset Y$.

Then, we have the following result.

Theorem 2. *Let $T : X \rightarrow Y$ be an operator between Banach spaces, and let $\mathbb{K} \subset X$ be a compact set. Then, by fixing $\epsilon > 0$ arbitrarily, there exists a Leray–Schauder transformer $\mathfrak{T}$ that approximates T with precision ϵ on $\mathbb{K}$, as follows:*

$$\|T(x) - \mathfrak{T}(x)\|_Y < \epsilon,$$

for all $x \in \mathbb{K}$.

Proof. The proof consists of an application of the results in [15,37]. As shown in Theorem 2.2 in [37], we can assume without loss of generality that T is uniformly continuous on a neighborhood V of $\mathbb{K}$, upon possibly replacing it by an approximation of T . Then, we can proceed as in the proof of Theorem 2.1 in [37] as follows. Since $\mathbb{K}$ is compact, we can construct a Leray–Schauder mapping $P : U \rightarrow X$ defined and continuous in an open neighborhood $U \subseteq V$ of $\mathbb{K}$, such that $\|x - P(x)\|_X < \delta$, where $\delta > 0$ is chosen in such a way that $\|T(z_1) - T(z_2)\|_Y < \frac{\epsilon}{3}$ whenever $\|z_1 - z_2\|_X < \delta$. From the construction of P , one sees also that $P(\mathbb{K}) \subset U$. Moreover, the image of P spans a finite dimensional subspace E of X , which is linearly homeomorphic to $\mathbb{R}^n$ for some n via a linear map $\varphi : E \rightarrow \mathbb{R}^n$.

Similarly, we can construct another Leray–Schauder mapping P' on the compact $T(P(\mathbb{K}))$, such that $\|y - P'(y)\|_Y < \frac{\epsilon}{3}$ for each $y \in T(P(\mathbb{K}))$. Moreover, P' can be continuously extended to the whole space Y . The image $P'(T(P(\mathbb{K})))$ is a finite dimensional subspace E' of Y that is linearly homeomorphic to $\mathbb{R}^m$ for some m via $\varphi' : E' \rightarrow \mathbb{R}^m$.

The map T induces a continuous mapping $T_{n,m} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ that can be approximated in the uniform norm over the compact $\varphi(P(\mathbb{K})) \subset \mathbb{R}^n$ by applying Corollary 3 in [15] or Theorem 1 in [23] (in conjunction with the traditional universal approximation results of functions by convolutional neural networks). In other words, we can approximate $T_{n,m}$ with arbitrary precision in the uniform norm with a polynomial-transformer architecture without positional encoding, as in [15], or with a traditional transformer architecture with positional encoding, as in [23]. Then, as in [37], we can show that the estimate $\|T(x) - \mathfrak{T}(x)\|_Y \leq \|T(x) - TP(x)\|_Y + \|TP(x) - P'TP(x)\|_Y + \|P'TP(x) - \mathfrak{T}(x)\|_Y < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$, where $\mathfrak{T} = \varphi'^{-1}W\varphi P : U \subset X \rightarrow E' \subset Y$.

□

We notice that the result of Theorem 2 gives a more general version of the results in [19], since their result is applicable to the case where X has the approximation property, a restriction that is not required in the present work. Here, the Leray–Schauder mappings play the role of the encoder used in [19].

To better elucidate the differences between Theorem 2 and the approximation results of [19], we provide a more detailed comparison.

- (i) The spaces X and Y in [19] are function spaces, defined over a compact subset of $\mathbb{R}^n$, while in our case's X and Y can be arbitrary Banach spaces.
- (ii) While X and Y in [19] are required to possess the approximation property, this is not needed in Theorem 2.
- (iii) No decoder is used in the present work, while the approximation found in [19] uses a decoder.
- (iv) The role of the encoder in [19] is played in Theorem 2 by the the Leray–Schauder mappings, which are explicitly and analytically defined (see Section 2).

4. Universal Approximation of Operators by Neural Integral Operators

In this section, we consider the universal approximation property of operators in Banach spaces, where the approximator is constructed through a neural integral operator in a suitable generalized sense.

Theorem 3. *Let X and Y be Banach spaces, let $T : X \rightarrow Y$ be an operator that is twice continuously Fréchet differentiable, and let $\mathbb{K} \subset X$ be a compact subset of X . For any $\epsilon > 0$, there exists an integral operator $U : X \rightarrow Y$ such that*

$$\|T(x) - U(x)\|_Y < \epsilon, \quad (10)$$

for all $x \in \mathbb{K}$.

Proof. We use the Taylor expansion for nonlinear operators found in Chapter 17 of Kantorovich–Akilov [27], where integration is intended as a Riemann integral over a Banach space as in [26]. See also [38,39]. Recall ([26,27]) that for $R : [x_0, \bar{x}] \subset X \rightarrow B(X, Y)$, where $B(X, Y)$ denotes the space of bounded linear operators from X to Y , the integral of R is defined (when it is well defined, e.g., when R is continuous) as

$$\int_{x_0}^{x_0+\Delta x} R(x)dx = \int_0^1 R(x_0 + t\Delta x)\Delta x dt = \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} R(x_0 + \tau_k \Delta x)\Delta x (t_{k+1} - t_k),$$

where $\tau_k \in [t_k, t_{k+1}]$ for all k . For a choice of $x_0 \in \mathbb{K}$ and for any $\bar{x} \in X$ such that T is twice continuously Fréchet differentiable over the interval $[x_0, \bar{x}]$ (i.e., the set of vectors $t\bar{x} + (1-t)x_0$ with $t \in [0, 1]$), we can write (Taylor's Theorem at degree 2)

$$T(\bar{x}) = T(x_0) + T'(x_0)(\bar{x} - x_0) + \int_{x_0}^{\bar{x}} T''(x)(\bar{x} - x, \cdot)dx, \quad (11)$$

where $\cdot$ is a placeholder for the entry of the bounded linear operator integrand. Since T'' is continuous, for each point $x \in X$, we can find a $\delta'_x > 0$ such that $\|T''(z) - T''(x)\|_{B^2(X,Y)} < \epsilon$ whenever $\|z - x\|_X < \delta'_x$, where $B^2(X, Y)$ denotes the space of bounded bilinear maps from X to Y with the norm given by $\|A\| = \sup_{\|h\|_X \leq 1, \|k\|_X \leq 1} \|A(h, k)\|_Y$. Then, for each x , define $\delta_x = \min\{\delta'_x, 1\}$. Since $\mathbb{K}$ is compact, we can cover it with finitely many balls $B(x_i, \delta_i)$, with $i = 1, \dots, n$, where δ_i is a shortened notation for δ_{x_i} . We set $A(t)(\cdot) := (T''(x(t)) - T''(x_i))((1-t)h, \cdot)$, with $x(t) = (1-t)h + x_i$, and we again use $\cdot$ to indicate where the input of the linear operator $A(t)$ for each $t \in [0, 1]$ is inserted. In correspondence with the balls $B(x_i, \delta_i)$, we define

$$T_i(\bar{x}) = T(x_i) + T'(x_i)(\bar{x} - x_i) + \int_{x_i}^{\bar{x}} T''(x_i)(\bar{x} - x, \cdot)dx. \quad (12)$$

Then, for any $x \in B(x_i, \delta_i)$, with $h = x - x_i$ and the parameterized interval $[x_i, x] \subset X$ given by $x(t) = (1-t)h + x_i$, one has

$$\begin{aligned} \left\| \int_0^1 A(t)dt \right\|_Y &\leq \int_0^1 \|A(t)\|_Y dt \\ &\leq \int_0^1 (1-t) \|h\|_X^2 \|T''(x(t)) - T''(x_i)\|_{B^2(X,Y)} dt \\ &< \epsilon. \end{aligned}$$

Therefore, for each $x \in B(x_i, \delta_i)$, we have

$$\begin{aligned} \|T(x) - T_i(x)\|_Y &= \left\| T(x_i) + T'(x_i)(\bar{x} - x_i) + \int_{x_i}^x T''(z)(x - z, \cdot) dz \right. \\ &\quad \left. - T(x_i) - T'(x_i)(\bar{x} - x_i) - \int_{x_i}^x T''(x_i)(x - z, \cdot) dz \right\|_Y \\ &= \left\| \int_{x_i}^x (T''(z) - T''(x_i))(x - z, \cdot) dz \right\|_Y \\ &= \left\| \int_0^1 (1-t)(T''((1-t)h + x_i))(h, h) dt \right\|_Y \\ &= \left\| \int_0^1 A(t) dt \right\|_Y \\ &< \epsilon. \end{aligned}$$

The space X is Hausdorff and paracompact, and we can therefore find a partition of unity $\{\eta_j\}_{j=1}^n \cup \{\eta_\infty\}$ subordinate to the open cover $\{B(x_i, \delta_i)\}_{i=1}^n \cup \{X - \mathbb{K}\}$ —see [40]—where we use the notation η_∞ to denote the element of the partition with support contained in $X - \mathbb{K}$. We now define the operator $U : X \rightarrow Y$ by patching together the operators $T_i : X \rightarrow Y$. In fact, for any $x \in X$, we set

$$U(x) = \sum_{i=1}^n \eta_i(x) T_i(x) + \eta_\infty(x) y_\infty,$$

where y_∞ is arbitrarily chosen. For any $x \in \mathbb{K}$, there exist $i_1, \dots, i_d \neq \infty$ such that $x \in B(x_i, \delta_i)$ for $i = i_1, \dots, i_d$ and $\sum_{i=i_1, \dots, i_d} \eta_i(x) = 1$, while $\eta_i(x) = 0$ for all $i \neq i_1, \dots, i_d$. Then, we have

$$\begin{aligned} \|T(x) - U(x)\| &= \left\| \sum_i \eta_i(x) T(x) - \sum_i \eta_i(x) T_i(x) \right\| \\ &= \left\| \sum_{i=i_1, \dots, i_d} \eta_i(x) T(x) - \sum_{i=i_1, \dots, i_d} \eta_i(x) T_i(x) \right\| \\ &\leq \sum_{i=i_1, \dots, i_d} \eta_i(x) \|T(x) - T_i(x)\| \\ &< \sum_{i=i_1, \dots, i_d} \eta_i(x) \epsilon \\ &= \epsilon, \end{aligned}$$

This completes the proof. $\square$

In the following, by the Gavuring neural integral operator $\mathfrak{T}(x)$, we mean the class of neural network architectures defined in Section 2.3.

Theorem 4. Let $T : X \rightarrow Y$ and $\mathbb{K}$ be as in Theorem 3. Then, for any choice of $\epsilon > 0$, there exists a Gavurin neural integral operator $\mathfrak{T}$ such that

$$\|T(x) - \mathfrak{T}(x)\|_Y < \epsilon,$$

for any $x \in \mathbb{K}$.

Proof. We apply Theorem 3 to approximate T via an integral operator U , where

$$U(y) = \sum_k \eta_k(y) T(y_k) + \sum_k \eta_k(y) T'(y_k)(y - y_k) + \sum_k \eta_k(y) \int_0^1 T''(y_k)((1-t)(y - y_k), y - y_k) dt,$$

in such a way that $\|T(y) - U(y)\|_Y < \frac{\epsilon}{2}$, where y_k are finitely many elements of $\mathbb{K}$, as in the proof of Theorem 3.

Let us consider the integral part of the operator U , namely

$$A(y) = \sum_k \eta_k(y) \int_0^1 T''(y_k)((1-t)(y - y_k), y - y_k) dt = \sum_k \eta_k(y) A_k(y),$$

where we set $A_k(y) := \int_0^1 K_k(t, y) dt$, with $K_k(t, y) := T''(y_k)((1-t)(y - y_k), y - y_k)$. We have that K_k is continuous with respect to t and y , and therefore uniformly continuous on $[0, 1] \times \mathbb{K}$. We select $\delta > 0$ such that $\|K_k(t_1, y_1) - K_k(t_2, y_2)\|_Y < \frac{\epsilon}{2}$ whenever $|t_1 - t_2| < \delta$ and $\|y_1 - y_2\|_X < \delta$, with $t_1, t_2 \in [0, 1]$ and $y_1, y_2 \in \mathbb{K}$. Given this choice of δ , we select a partition of points $0 = t_0, t_1, \dots, t_{d-1}, t_d = 1$ such that $|t_i - t_{i+1}| < \delta$ for all $i = 0, 1, \dots, d-1$. Using Theorem 2.1 in [37], we can find a neural network $\hat{K}_{ij}$, consisting of Leray–Schauder mappings and a deep neural network, such that $\|K_i(t_j)(y) - \hat{K}_{ij}(y)\|_Y < \frac{\epsilon}{2}$ for all $y \in \mathbb{K}$. Let us now consider the open covering of $[0, 1]$ with $B_\delta(t_j)$ open balls centered at t_j , with $j = 0, \dots, d$, and radius δ . Let α_k denote a partition of unity subordinate to $\{B_\delta(t_j)\}$. One can directly verify that the operator $\hat{K}_i(t) := \sum_k \alpha_k(t) \hat{K}_{ik}$ satisfies the property that $\|K_i(t)(y) - \hat{K}_i(t)(y)\|_Y < \frac{\epsilon}{2}$ whenever y is in $\mathbb{K}$. Since this construction can be performed for each i , it follows that the integral part of the operator U can be approximated by an integral operator with a kernel consisting of Leray–Schauder mappings and (deep) neural networks.

Another application of Theorem 2.1 in [37] allows us to approximate $\sum_k \eta_k(y) T'(y_k)(y - y_k)$ with a deep neural network (and Leray–Schauder mappings) with precision $\frac{\epsilon}{2}$ over $\mathbb{K}$. Putting both approximations together completes the proof. $\square$

5. Further Perspectives

We conclude this article with some observations for further investigation.

We would like to mention that the results of Theorem 1 approximate integral operators between Hölder spaces, where the need for Hölder norms arises from the use of numerical quadrature to approximate integrals. Therefore, one might ask under what conditions transformers are universal approximators of integral operators between L^p spaces. This question is yet to be answered.

Theorem 2 gives an approximation result for transformers in complete generality—note that it refers to general Banach spaces, rather than some specific Banach space of functions. Once a norm is fixed, the Leray–Schauder mapping has an explicit form that is analytical, and therefore, it seems that the new architecture is completely transparent. In general, however, it is useful to learn the Leray–Schauder mappings as well to increase the applicability of the model. This is done, for example, in [36], where the resulting model is tested (in the L^p -space) on some integral equations and PDEs. The theoretical guarantees are also studied. It would be interesting to investigate extensions of these results in relation to the work in this article. The main difficulty in extending the present work to the case of L^p -spaces, for example, is that there is no bound on the derivatives that can control the error associated to an integration scheme. Therefore, an analogous version of Lemma 1 would fail. It is possible to attempt similar approaches in the Sobolev spaces, as cubature

formulas can control the integration errors. However, we will leave this interesting line of work to future research.

Lastly, the results of Section 4 show that mappings between general Banach spaces can be approximated through a generalized version of the neural integral operator based on the Gavurin integral. While this is an interesting theoretical development, it is still unclear whether removing the use of the Gavurin integral would still allow for such a general approximation result.

Author Contributions: E.Z. conceived the idea and initiated the project. E.Z. and M.B. jointly developed the proofs and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Chen, T.; Chen, H. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans. Neural Netw.* **1995**, *6*, 911–917. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lu, L.; Jin, P.; Pang, G.; Zhang, Z.; Karniadakis, G.E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.* **2021**, *3*, 218–229. [\[CrossRef\]](#)
- Zappala, E.; Fonseca, A.H.d.O.; Moberly, A.H.; Higley, M.J.; Abdallah, C.; Cardin, J.A.; van Dijk, D. Neural integro-differential equations. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023. Volume 37, pp. 11104–11112.
- Zappala, E.; Fonseca, A.H.d.O.; Caro, J.O.; Moberly, A.H.; Higley, M.J.; Cardin, J.; Dijk, D.V. Learning integral operators via neural integral equations. *Nat. Mach. Intell.* **2024**, *6*, 1046–1062. [\[CrossRef\]](#)
- Zappala, E.; Fonseca, A.H.d.O.; Caro, J.O.; van Dijk, D. Neural Integral Equations. *arXiv* **2022**, arXiv:2209.15190.
- Torlai, G.; Mazzola, G.; Carrasquilla, J.; Troyer, M.; Melko, R.; Carleo, G. Neural-network quantum state tomography. *Nat. Phys.* **2018**, *14*, 447–450. [\[CrossRef\]](#)
- Mitarai, K.; Negoro, M.; Kitagawa, M.; Fujii, K. Quantum circuit learning. *Phys. Rev. A* **2018**, *98*, 032309. [\[CrossRef\]](#)
- Hwang, R.; Lee, J.Y.; Shin, J.Y.; Hwang, H.J. Solving PDE-Constrained Control Problems Using Operator Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online Conference, 22 February–1 March 2022; Volume 36, pp. 4504–4512. [\[CrossRef\]](#)
- Li, Z.; Zheng, H.; Kovachki, N.; Jin, D.; Chen, H.; Liu, B.; Azizzadenesheli, K.; Anandkumar, A. Physics-Informed Neural Operator for Learning Partial Differential Equations. *ACM/IMS J. Data Sci.* **2024**, *1*, 1–27. [\[CrossRef\]](#)
- Hashimoto, K.; Hu, H.Y.; You, Y.Z. Neural ordinary differential equation and holographic quantum chromodynamics. *Mach. Learn. Sci. Technol.* **2021**, *2*, 035011. [\[CrossRef\]](#)
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
- Guo, R.; Cao, S. Transformer Meets Boundary Value Inverse Problem. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
- Geneva, N.; Zabarar, N. Transformers for modeling physical systems. *Neural Netw.* **2022**, *146*, 272–289. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cao, S. Choose a transformer: Fourier or galerkin. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24924–24940.
- Fang, Z.; Ouyang, Y.; Zhou, D.X.; Cheng, G. Attention enables zero approximation error. *arXiv* **2022**, arXiv:2202.12166.
- Yun, C.; Bhojanapalli, S.; Rawat, A.S.; Reddi, S.J.; Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
- Takakura, S.; Suzuki, T. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 33416–33447.
- Jiao, Y.; Lai, Y.; Sun, D.; Wang, Y.; Yan, B. Approximation bounds for transformer networks with application to regression. *arXiv* **2025**, arXiv:2504.12175.
- Shih, B.; Peyvan, A.; Zhang, Z.; Karniadakis, G.E. Transformers as neural operators for solutions of differential equations with finite regularity. *Comput. Methods Appl. Mech. Eng.* **2025**, *434*, 117560. [\[CrossRef\]](#)

20. Lanthaler, S.; Mishra, S.; Karniadakis, G.E. Error estimates for deeponets: A deep learning framework in infinite dimensions. *Trans. Math. Its Appl.* **2022**, *6*, tnac001. [\[CrossRef\]](#)
21. Deng, B.; Shin, Y.; Lu, L.; Zhang, Z.; Karniadakis, G.E. Approximation rates of DeepONets for learning operators arising from advection–diffusion equations. *Neural Netw.* **2022**, *153*, 411–426. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Lu, L.; Jin, P.; Karniadakis, G.E. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv* **2019**, arXiv:1910.03193.
23. Cordonnier, J.B.; Loukas, A.; Jaggi, M. On the relationship between self-attention and convolutional layers. *arXiv* **2019**, arXiv:1911.03584.
24. Precup, R. *Theorems of Leray-Schauder Type and Applications*; CRC Press: Boca Raton, FL, USA, 2002.
25. Krasnosel'skii, Y.P. *Topological Methods in the Theory of Nonlinear Integral Equations*; Pergamon Press: Oxford, UK, 1964.
26. Gavurin, M. Über die Stieltjessche Integration abstrakter Funktionen. *Fundam. Math.* **1936**, *27*, 254–265. [\[CrossRef\]](#)
27. Kantorovich, L.V.; Akilov, G.P. *Functional Analysis*; Elsevier: Amsterdam, The Netherlands, 2016.
28. Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R.R.; Smola, A.J. Deep sets. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
29. Zhou, A.; Finn, C.; Harrison, J. Universal neural functionals. *arXiv* **2024**, arXiv:2402.05232.
30. Young, D.M.; Gregory, R.T. *A Survey of Numerical Mathematics*; Courier Corporation: Lowell, MA, USA, 1988; Volume 1.
31. De Oliveira Fonseca, A.H.; Zappala, E.; Caro, J.O.; Van Dijk, D. Continuous spatiotemporal transformer. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; ICML/23.
32. Amari, S.I. Homogeneous nets of neuron-like elements. *Biol. Cybern.* **1975**, *17*, 211–220. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Amari, S.I. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* **1977**, *27*, 77–87. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Wilson, H.R.; Cowan, J.D. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik* **1973**, *13*, 55–80. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Pötzsche, C. Urysohn and Hammerstein operators on Hölder spaces. *Analysis* **2022**, *42*, 205–240. [\[CrossRef\]](#)
36. Zappala, E. Leray-Schauder Mappings for Operator Learning. *arXiv* **2024**, arXiv:2410.01746.
37. Zappala, E. Projection Methods for Operator Learning and Universal Approximation. *arXiv* **2024**, arXiv:2406.12264.
38. Gordon, R. Riemann integration in Banach spaces. *Rocky Mt. J. Math.* **1991**, *21*, 923–949. [\[CrossRef\]](#)
39. Graves, L.M. Riemann integration and Taylor's theorem in general analysis. *Trans. Am. Math. Soc.* **1927**, *29*, 163–177. [\[CrossRef\]](#)
40. Dugundji, J. Allyn and Bacon Series in Advanced Mathematics. *Topology* **1966**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.